Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 30 (2014) 39 – 49

Procedia Computer Science

1st International Conference on Data Science , ICDS 2014

Economic Regionalization Based On Order-Preserving Submatrix

Gege Zhang, Weixing Zhou, Xiaohui Hu, Yun Xue, Xiaosheng Wu, Tiechen Li, Zhenglin Liao, Hua Xiao

School of Physics and Telecommunications,South China Normal University,Guangdong 510006,China

Abstract

In the paper, We compared the classical K-means clustering algorithm, the fuzzy clustering with the Order-Preserving Submatrix(OPSM) biclustering algorithm on the dataset of regional fiscal revenue which is collected from the National Statistical Yearbook. The experimental results proves that the OPSM biclustering algorithm could get more interesting results than the classical K-means algorithm and the fuzzy clustering algorithm, which shows more detailed information than the latter either.

Keyword: K-means algorithm, Fuzzy clustering, Order-Preserving Submatrix,biclustering, fiscal revenue

1. Introduction

Fiscal revenue is the total capital for the government to fulfill its function such as the implementation of public policy, the provision of public goods and services. As a comprehensive developing country, the regions in China are different in economy, population, resources and environment ,etc, thus cause difference in regional financial income. This paper clusters the regions into different classes by analyzing 19 public index of the area' fiscal revenue in 10 years

In the development of regional economic, scientificly dividing regions by the high similarity is a good way to take advantage of the regional feature to promote the economic development. Generally there are two kinds of methods to solve the problem: traditional classification and numerical classification[1].The traditional classification is usually based on experience and relevant expertise, which is a qualitative classification. Although it can achieve certain results, they are still ambiguous to some degree, which makes difficult in more detailed description about the difference and relations among the objects. Sometimes the researcher's subjective intention affected the classified objectivity. The numerical classification(mainly cluster analysis) can avoid the shortcomings of traditional classification of subjectivity and arbitrariness in a certain extent. However, traditional cluster analysis can only find the global information, not the local information[6]. By using different

doi:10.1016/j.procs.2014.05.379

* Corresponding author. Tel.: 13503064048; fax:+39310066.
E-mail address: 1213268188@qq.com, zggscun@163.com, zhouwx@scnu.edu.cn, 1085206157@qq.com, xueyun@scnu.edu.cn, 2454922794@qq.com, 352385567@qq.com, 739613590@qq.com

methods and sample data, we will have different final classification results. Therefore we need to combine the subjective judgments with the objective facts [16]to obtain a more rational analysis[6].

In the paper the data of regional fiscal revenue is collected from the National Statistical Yearbook. The characteristics of such a number of years' revenue are analyzed in both time and spatial distribution[1]. The biclustering method is applied to discuss the regional fiscal revenue to findthe local correlation and more detailed information.

2.Clustering

Here we use two different methods, one is the classical K-means clustering method, another is fuzzy clustering, which are introduced below.

2.1.Division Clustering

Division Clustering uses the given number k to create an initial partition, then use an iterative relocation, trying to improve the division by moving the object between the division. A general rule is: the distance between objects in the same class is as close as possible , and as far as possible between different class . In order to reach the global optimum, partition cluster requires the enumeration based on all possible partition. Partition method used in this paper is the K-means algorithm.

2.2. The K-means algorithm

K-means algorithm is the centroid based algorithm. The K-means algorithm divide n objects into K clusters (K is the given number), with high similarity in the same cluster and high dissimilarity in different cluster. Sim ilarity is calculated by the average value of the objects in a cluster (center or gravity of a cluster) . The specifi c process of K-means clustering algorithm is described as follows[9].

Step1 Select k objects as the initial cluster center from the data set as C_1, C_2, …, C_k;

Step2 The rest objects are assigned to the cluster whose distance is the shortest. The most similar refers to the minimum distance. For each point V_i, Find a centroid C_j, if there exists the minimum distance between them, then assigned V_i to the group j;

Step3 After all the points are assigned to clusters, the centroid C_j is recalculated for each group;

Step4 Goto Step2, otherwise stop until data will not change.

2.3.Fuzzy Clustering

Fuzzy clustering is a clustering algorithm which each data point belongs to a cluster with the degree of member ship[7].The key point of the algorithm is to find the membership function. Here we will provide a method for det ermining the membership function by the statistical analysis method[2].

1) The sample data is ordered in ascending, then a sample interval $[x_{(1)}, x_{(n)}]$ is set, the nodes are determined isometricly,ie. dividing 5 isometric intervals, each interval is $\frac{x_{(n)}-x_{(1)}}{5}$

Set membership function:

Dropping type[8]

$$U_A(x) = \begin{cases} 1 & x \le a \\ F_1(x) & a < x \le b \\ 0 & b < x \end{cases} \quad (1)$$

Intermediate type[8]

$$U_B(x) = \begin{cases} 0 & x \le a \\ F_2(x) & a < x \le b \\ F_3(x) & c < x \le d \\ 0 & d < x \end{cases} \quad (2)$$

Rising type[8]

$$U_C(x) = \begin{cases} 0 & x \le c \\ F_4(x) & c < x \le d \\ 1 & d < x \end{cases} \quad (3)$$

Where a, b, c, d is determined by the way of equidistant nodes, $F_2(x)$, $F_4(x)$ are increasing functions, $F_1(x)$, $F_3(x)$ are decreasing functions. First to obtain $F_2(x)$, let s be the number of samples in the range of $\left(a, \frac{a+b}{2}\right)$, t be the number of samples in the range of $\left(\frac{a+b}{2}, b\right)$, thus the parabola across the 3 point s $(a, 0), \left(\frac{a+b}{2}, \frac{s}{s+t}\right), (b, 1)$ is confirmed. It satisfies the following equation

$$\begin{cases} a^2 x_1 + a x_2 + x_3 = 0 \\ \left(\frac{a+b}{2}\right)^2 x_1 + \frac{a+b}{2} x_2 + x_3 = \frac{s}{s+t} \\ b^2 x_1 + b x_2 + x_3 = 1 \end{cases} \quad (4)$$

Solutions of three element linear equation group:

$$F_2(x) = \frac{1}{h}(p x^2 + q x + r), a < x \le b$$

$$h = (b-a)^2(s+t)$$

$$p = 2(t-s)$$

$$q = a(s-3t) + b(3s-t)$$

$$r = a^2(s+t) + ab(t-3s)$$

When $s = t$ then $F_2(x) = \frac{x-a}{b-a}$

When $s > t$ then $F_2(x)$ In $(a, b]$ is convex

When $s < t$ then $F_2(x)$ In $(a, b]$ is concave

As mentioned above, $F_1(x)$ can be acquired. $F_1(x)$ and $F_2(x)$ are symmetry of $\frac{1}{2}$

Then

$$F_1(x) = 1 - F_2(x) = -\frac{1}{h}(p x^2 + q x + r - h) \quad a < x \le b$$

$$F_4(x) = \frac{1}{h_1}(p_1 x^2 + q_1 x + r_1) \quad c < x \le d$$

$$F_3(x) = -\frac{1}{h_1}(p_1 x^2 + q_1 x + r_1 - h_1) \quad c < x \le d$$

The calculation of h_1, p_1, q_1, r_1 is the same as h, p, q, r, with a, b, s, t being replaced by c, d, s_1, t_1[8]

2) Once the membership function is determined, calculate the distance $d(i, j)$ between the object and the clustering center, if it is less than a certain threshold, or the variance from the last value of the membership function is less than a threshold, then the algorithm stops.

3.Biclustering

Order-preserving Submatrix [8](OPSM) has been introduced and accepted as abiologically meaningful

cluster model. An OPSM is,essentially a pattern-based subspace cluster, a subset of rows and columns in a data matrix, as shown in Fig.1. An OPSM cluster may arise when the expression levels of the coregulated genes rise and fall synchronously in response to a sequence of environment stimuli. To discover significant OPSMs from a given data matrix plays an essential role in inferring gene regulatory networks[12].

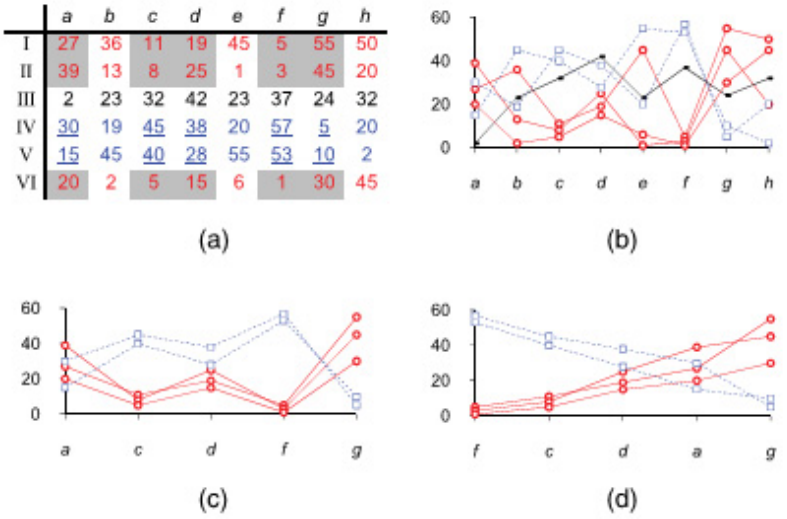


Fig.1.OPSM and GOPSM. The raw gene expression data matrix in (a) Exhibits no obvious pattern when plotted in(b).(c) Shows an GOPSM consisting of two OPSMs corresponding to the submatrices with shaded entries and underline dentries in the data matrix respectively. (d) Shows a permutation of columns of the GOPSM, under which the row sequences are in either strictly ascending or descending order.(a) Data matrix.(b) Data matrix plotted.(c) GOPSM consisting of two OPSM.(d) GOPSM rearranged.[12].

The OPSM cluster model focuses on the relative order of columns rather than the uniformity of actual values in data matrices[10].By sorting the row vectors and replacing the entries with their corresponding column labels, the data matrix can be transformed into a sequence database[13],and OPSM mining is reduced to a special case of the sequential pattern mining problem with some distinctive properties[17][18].In particular, the sequence data base is extremely dense since each column label appears exactly once (assuming no missing values) in each sequence. A sequential pattern uniquely specifies an OPSM with all the supporting sequences as the rows. The number of supporting sequences is the support count, or simply support,for the pattern[12]

The OPSM problem[8] is to discover those statistically significant OPSMs from the data matrix .The problem is closely related to sequential pattern mining. Conventional sequential pattern mining was [8]motivated and introduced in the context of transaction databases, where a sequence is an ordered list of item sets[15]. A common subsequence with support (number of sequences containing the subsequence) beyond a minimum support threshold, min_sup,is called a frequent sequential pattern. The sequential pattern mining problem is to find the complete set off requent sequential patterns with respect to min_sup[14].

OPSM mining can be reduced to a special case of the sequential pattern mining problem[12]. If we sort each row in the datamatrix D in ascending order, and replace the entries with their corresponding column labels, then D is transformed into a sequence database, as shown in Fig.3b. Each sequential pattern uniquely specifies an OPSM, where the pattern specifies the columns and the supporting sequences specify the rows.

In this paper,we found the OPSMs from the matrix whose rows are 31 areas and the columns are years from 2003 to 2012. We form a matrix by each index, then find the OPSM for each matrix(including the OPSM of max colum). Finally, we figure out the areas which had the most common indice and cluster them together.

4.Experiment and results

4.1. Data set

We collect the revenue data of 31 Chinese provinces (including municipalities and autonomous regions) from 2003 to 2012.Then select the 19 indexes from it (Table 1). The data for each year are springing from the State Statistical Yearbook[19]. The 10 years' data of each region is normalized by the formula $(x_i - x_{min})/(x_{max} - x_{min})$, so that the data are mapped in the range (0,1).

Table 1 Revenue indice of each region

No. Index	No. Index	No. Index
1 Domestic VAT	8 Stamp Duty	15 Administrative fee income
2 Business tax	9 Urban Land Use Tax	16 Forfeiture income
3 Corporate Income Tax	10 LAT	17 State-owned capital operating income
4 Personal Income Tax	11 Travel tax	18 Income paid use of state resources
5 Resource tax	12 Arable land occupation tax	19 Other income
6 Urban maintenance and construction tax	13 Deed	
7 Property Tax	14 Special revenue	

4.2. The results of K-means clustering method

This paper used the classical K-means algorithm to cluster the regions，the results are stated as follows[1]:

Table 2 The K-means clustering results

The first class	The second class	The third class	The fourth class	The fifth class
Beijing	Zhejiang	Tianjin	Liaoning Province	Tibet
Shanghai	Guangdong	Hebei	Jilin	Inner Mongolia
	Fujian Province	Jiangsu Province	Heilongjiang	
	Shanxi	Anhui	Hubei	
	Henan	Jiangxi Province	Sichuan Province	
		Shandong	Guizhou	
		Hunan	Yunnan	
		Guangxi	Shanxi Province	
		Hainan	Gansu Province	
		Chongqing	Qinghai	
		Ningxia	Xinjiang	

The first class consists of Beijing, Shanghai which are two municipalities. Beijing and Shanghai share of a large proportion in domestic value-added tax, business tax, corporate income tax, property tax. This is because Beijing and Shanghai are the region economic and trade center, which have adequate sources of revenue, with complete services in transportation, finance and insurance, post and telecommunications, culture and sports.. The business tax, corporate income tax and property taxes account for 75.32% (Beijing) and 76.53% (Shanghai) of the average of domestic VAT. At the same time, the average urban land using tax in Beijing and

Shanghai were 1.057 billion and 1.657 billion（￥) respectively, which are the average of the whole nation. This is the style of municipality cities, ,which are characterized by limited land.

The second class contains Zhejiang, Guangdong, Fujian, Shanxi, Henan. Zhejiang, Guangdong, Fujian are the coastal provinces, and the gross taxation is higher than the national average taxation. so they are clustered. Although the domestic VAT and business tax of Zhejiang and Guangdong are higher than those of Fujian, Shanxi and Henan, the average growth rate of domestic value-added tax of Shanxi, Zhejiang, Fujian, Henan and Guangdong is 14.2%, 14%, 15%, 14%, 15% ,respectively, and the average growth rate of sales tax of Shanxi, Zhejiang, Fujian, Henan and Guangdong is 22%, 19%, 22%, 23%, 19%, which shows a stable similiar growth trend.

The third class includes 11 regions. The domestic VAT, business tax, corporate income tax and personal income tax of Jiangsu and Shandong are statistically a little higher than the average of the third category. Other areas did not differ much with each other. From 2003 to 2012, the average growth rates of sales tax are23%(Tianjin), 26%(Hebei), 26%(Jiangsu), 29%(Anhui), 27%(Jiangxi), 25% (Shandong), 24%(Hunan), 24%(Guangxi), 27%(Hainan), 26%(Chongqing) , 27%(Ningxia). Therefore, they are classified into a class.

The fourth category consists of 11 regions. These regions have no particular geographical advantage, which have large labor force, relatively large number of migrant workers. The average contribution to the country's fiscal revenue is 1% ~ 2% from 2003 to 2012. The domestic value-added tax, business tax, corporate income tax, personal income tax grow up slowly in these regions.

The fifth category concludes Tibet and Inner Mongolia. The contribution is less than 0.5% in recent 10 years. They are short of labor force and need the national support.

4.3. Fuzzy Clustering Analysis

After the fuzzy processing the normalized data with the given membership function the clustering results are depicted as follows:

Table 3 Results of Fuzzy Clustering

The first class	The second class	The third class	The fourth class	The fifth class
Beijing	Guangdong	Tianjin	Liaoning	Xizang
Shanghai	Jiangsu	Hebei	Jilin	Shanxi
Zhejiang	Shandong	Henan	Heilongjiang	Inner Mongolia
	Fujian	Anhui	Sichuan	
	Shanxi	Jiangxi	Guizhou	
		Hunan	Yunnan	
		Hubei	Gansu	
		Guangxi	Qinghai	
		Hainan	Xinjiang	
		Chongqing		
		Ningxia		

Compared with Table 2, the clustering results of Table 3 are changed a little Zhejiang is divided into the first class of Table 3 which is in the second class of Table 2. According to the method we used in the fuzzy clustering, after equally divided the membership ranges, Zhejiang shows the same character with Beijing and Shanghai. There appears the same pattern among Xizang, Shanxi and Inner Mongolia(The fifth class). Jiangsu ,Shandong are divided into the second class of Table 3 which are in the third class of Table 2. Henan is divided into the third class of Table 3 which is in the second class of Table 2. Hubei is divided into the third

class of Table 3 which belongs to the fourth class of Table 2. However the core areas of each class do not change as a whole.

4.4. The Analysis of Biclustering Results

After find the OPSMs from the 19 indexes of the 31 areas in 10 years, we cluster the area who have the same OPSM model. The results are shown in following Figure.

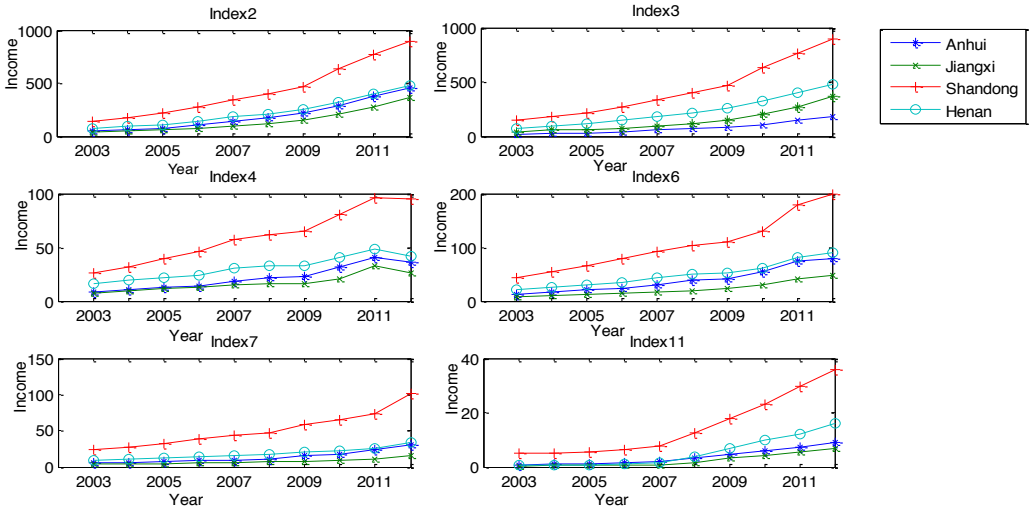
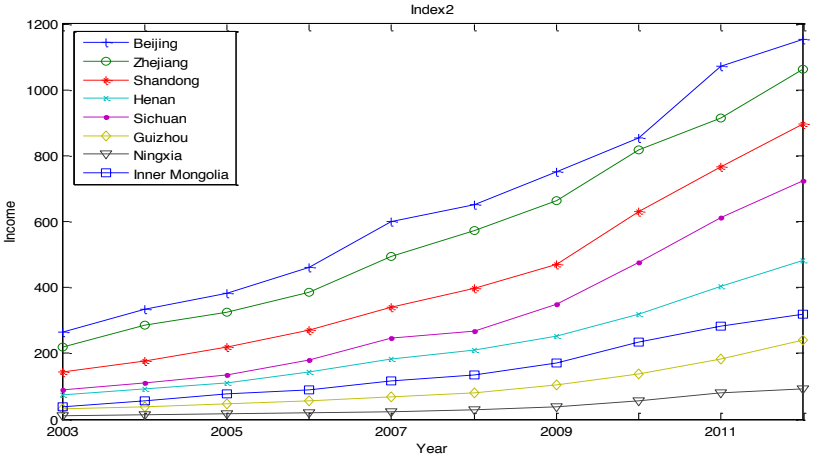
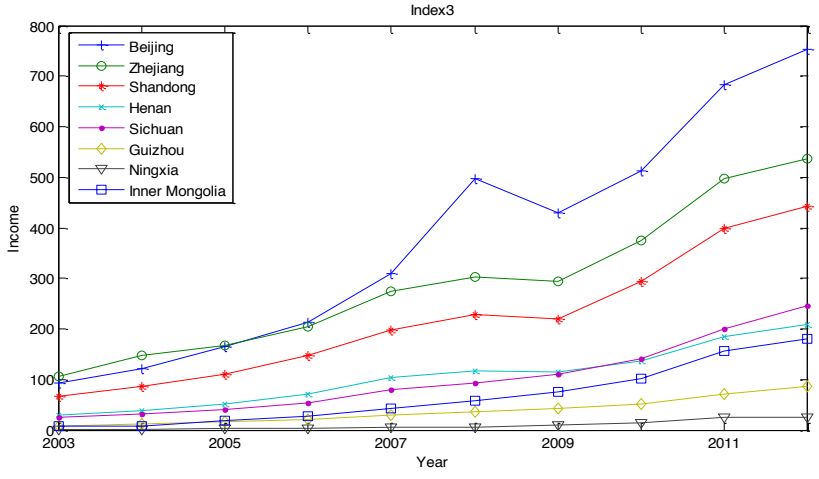


Fig. 2 OPSM 1

Figure 2 shows that Anhui, Jiangxi, Shandong, Henan have the similar trends of tax income in the index 2, index3, index4, index6, index7, index11. It is obvious that the actual value of each area in a year are differencet, but the trends are similar. In other words, the 4 areas' tax income of index2, index3, index4, index6, index7, index11 will rise or fall in the same year.


Index2
Beijing
Zhejiang
Shandong
Henan
Sichuan
Guizhou
Ningxia
Inner Mongolia
Income
Year
0
200
400
600
800
1000
1200
2003
2005
2007
2009
2011


Index3
Beijing
Zhejiang
Shandong
Henan
Sichuan
Guizhou
Ningxia
Inner Mongolia
Income
Year
0
100
200
300
400
500
600
700
800
2003
2005
2007
2009
2011

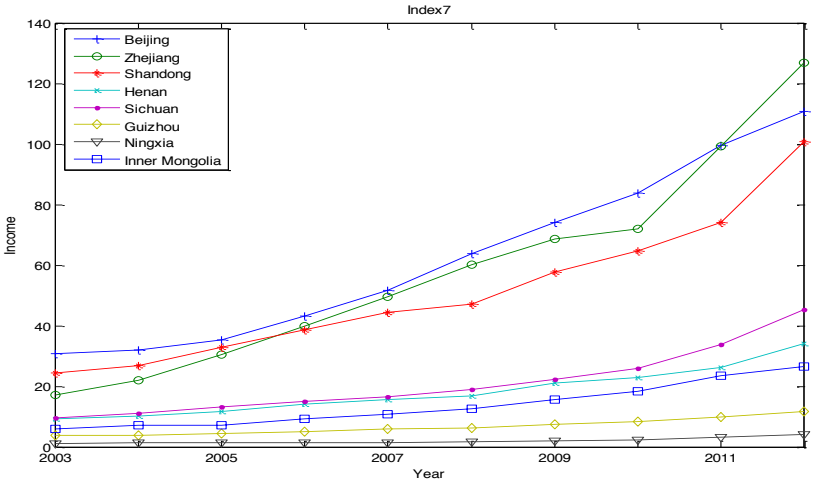


Fig. 3 OPSM 2

In Figure 3, Beijing, Zhejiang, Shandong share the similar trend in index2, index3, index7 in the continuous time. Taking index2 as an example, the tax income of Beijing from 2003 to 2012 is nearly 22 times of Ningxia's tax income.If we applies the method of K-means clustering, the two areas won't get together. While the similarity of data trend could be found among the areas.

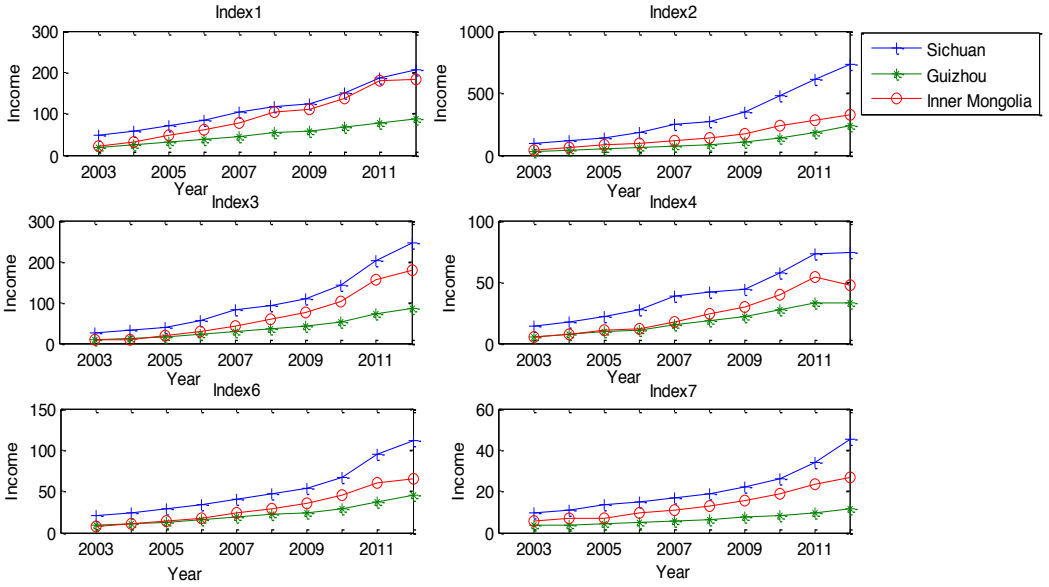


Fig. 4 OPSM 3

Figure4 shows Sichuan,Guizhou,Inner Mongolia has the similar trend of index1, index2, index3,

index4,index6,index7.So they can be clustered together.

5.Conclusion

In summary, this paper applies traditional clustering algorithms as well as biclustering algorithm to analysis t he regions financial data, and obtains some useful results [4].

The results of the K-means clustering and fuzzy clustering are of a little difference. Beijing, Shanghai are th e economic center directly under the central government. Their domestic value-added tax, business tax, enterpr ise income tax, personal income tax, property tax base account for a the high proportion. At the same time, Zhe jiang as a big coastal province has some similar characteristic with Beijing and Shanghai. In addition, the clust ering results prove that the fiscal income of coastal areas are not similiar, such as Guangdong and Guangxi. T here are a large difference between the two regional financial income, therefore they belong to different classes.

The method of finding the OPSM is to cluster the area depending on the trend not the real value[10]. It can hel p us to get more explicit information. On one hand, the regions not belonging to the same class under the traditi onal clustering algorithm could gather together under biclustering algorithm. For example, in figure4 Ningxia, I nner Mongolia are in the same cluster with Beijing under some indexes , which shows that although there is a d isparity among the integral developing levels of different provinces, it may appear similar trend in some attribut es. On the other hand, the same areas appear in different clusters, such as Sichuan, Guizhou, Inner Mongolia (figure3) also emerge in figure 4, breaking through the limitations of the traditional clustering algorithm that a n area only belongs to a cluster.

Acknowledgements

The authors thank gratefully for the colleagues who have been concerned with the work and have provided much more powerfully technical supports. The work is sup-ported by Guangdong Science and Technology Dep artment under Grant No.2009B090300336,No.2012B091100349; Guangdong Economy & Trade Committee un der Grant No. GDEID2010IS034; Guangzhou Yuexiu District science and Technology Bureau un-der Grant No 2012-GX-004; National Natural Science Foundation of China (Grant No:71102146, No.3100958); Science and Technology Bureau of Guangzhou under Grant No. 2011J4300046.

Reference

1.Jing Zhao,Qin Ma,Yuquan Cui (2012). A comparative study of macroeconomic divisions: double clustering algorithm application. Journal of ShanDong University,47(9),71-77

2.Shiqing Wang,Lin Qiu,Zhiliang Wang,Xiaojun Han (2001).Statistical Analysis to Determine the Membership Function. Journal of North China Institute of Water Conservancy and Hydripower. 23(1),68-71.

3.Yanwen Zhang (2006).Regional Economic Disparities Based Spatial Clustering Analysis (2006). Journal of Economic Geography. 7,557-560.

4.Chunli Du,Jinhua Cheng.Circulating Levels of Economic Development in All Regions of Soft Clustering Analysis (2009). Journal of Operation Research and Management .6,116-122.

5.LI Cai,Hong Guo.An Improved Algorithm for Clustering Gene Expression Data Pairs (2010). Journal of Fu Zhou University. 38(1),41-47.

6.Min Zhang,Wenhang Ge.Research and Development of BiClustering. Journal of Review and Comment.

7.Hongxing Li,Peizhuang Wang.(1994).Fuzzy Math.BeiJing:Defense Industry Press.

8. A.Ben-Dor,B.Chor,R.Karp,andZ.Yakhini,"Discovering Loca Structure in Gene Expression Data:The Order-Preserving Submatrix Problem", J.Computational Biology, vol.10,no.3/4, pp.373-384,2003

9.Changan Yuan.(2009).Principles and Applications of Data Mining.BeiJing:Electronic Indutry Press.

10.S.Madeira and A.Oliveira,"Biclustering Algorithms for Biological Data Analysis:A Survey", IEEE/ACM Trans.Computational Biology and Bioinformatics, vol.1,no.1,pp.24-45,Jan.-Mar.2004.

11.R.Agrawal and R.Srikant, "Mining Sequential Patterns", Proc 11th Intl Conf. Data Eng.(ICDE) 1995.

12.Byron J.Gao,Obi L.Griffith,Martin Ester,Hui Xiong,Qiang Zhao,and Steven J.M.Jones,"On the Deep Order-Preserving Submatrix Problem:A Best Effort Approach ",IEEE Transactions On Knowledge And Data Engineering, vol.24,no.2,February 2012.

13.C.C.Aggarwal,J.L.Wolf,P.S.Yu,C.Procopiuc,and J.S.Park,"Fast Algorithms for Projected Clustering," SIGMOD Record,vol.28,no.2,pp.61-72,1999.

14.R.Agrawal ,J.Gehrke,D.Gunopulos,and P.Raghavan,"Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications,"SIGMOD Record ,vol.27,no.2,pp.94-105,1998.

15.H-P.Kriegel,P.Kroger, and A.Zimek,"Clustering High-Dimensional Data:A Survey on Subspace Clustering , Pattern-based Clustering, and Correlation Clustering,"ACM Trans.Knowledge Discovery Data,vol.3,no.1,pp.1-58,2009.

16. Yang Juan,Wang Chang-quan,Li Bing, Li Qi-quan, Song Wei-ping . Self-organized Competing Neural Network and Its Application in Compartmental izing Social Economy Areas. Journal of Southwest China Normal University (Natural Science) Vol . 32 No. 4 Aug. 2007

17. L. Jensenetal. , Arrayprospector:A Web Resource of Functional Associations Inferred from Microarray Expression Data, Nucleic Acids Research, vol. 32 ,pp.W445-W448,2004.

18.J.B. MacQueen,"Some Methods for Classifiation and Analysis of Multivariate Observations", Proc .Fifth Berkeley Symp. Math. Statistics and Probability, 1967.

19. The people's Republic of China National Bureau of Statistics. Chinese statistical yearbook(2012)[M],Beijing: National Statistics Press,2012